DOCUMENT RESUME

ED 200 628 TM 810 213

AUTHOR Cason, Gerald J.
TITLE Guidelines for Objective Testing of Students.
INSTITUTION Arkansas Univ. for Medical Sciences, Little Rock.
PUB DATE Mar 78
NOTE 62p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Cognitive Tests; *Objective Tests; *Scoring; *Test Coaching; *Test Construction; *Testing; *Test Interpretation

ABSTRACT

Guidelines developed specifically for cases in which objectively scored written tests are to be used are described. When objective cognitive tests are properly built and used, they can serve four educationally important functions. These tests provide the instructor information on what a student can and cannot do. The instructor needs this information both to select and/or design further instruction and to decide if and/or when a student is qualified for advancement to the next unit, block, course, or level of instruction. Well constructed and properly used objective cognitive tests provide the student a guide to study and confirmation of what he is able to do. Sections provide guidelines for: (1) constructing the test; (2) informing the students about the test; (3) giving the test; (4) scoring and analyzing the test; and (5) reporting results of the test. In addition to the guidelines, some explanatory material and supporting rationale are provided. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED200628

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

GUIDELINES FOR OBJECTIVE TESTING OF STUDENTS

Gerald J. Cason, Ph.D.
Office of
Educational Development

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. J. Cason

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

March 1978

University of Arkansas for Medical Sciences
Little Rock, Arkansas

TM 810 213

PREFACE

I have written this paper as a practical guide and reference on objective testing. I expect that most frequently it will be used when a faculty member is engaged in test preparation or the part of course planning concerned with the assessment of student performance. Because of the heavy demands upon faculty time, I doubt that many instructors will find the time to systematically study and learn all the information presented here. For that reason, I have organized the material using numerous paragraph headings and extensive indentations. These features combined with the detailed table of contents should make it relatively easy to locate information when it is needed.

I would strongly recommend to anyone directly concerned with testing that they at least scan the contents of this entire paper. However, I recognize that some of the information is more relevant to some audiences than others. The three general groups of readers who should be familiar with parts of this paper are educational administrators, test constructors, and item writers.

In the general category of educational administrators, I included Department Chairmen, Program Directors, and Assistant Deans who are administratively responsible for the overall quality of the educational practices within a collection of courses. The information most relevant to this group is indicated by the items having two or three asterisks in the table of contents.

The sections of the paper of most relevance to those concerned almost solely with writing items are indicated by one or three asterisks in the table of contents.

Within the general group that I call test constructors, I include those individuals with direct overall responsibility for the construction, administration, scoring, analysis and interpretation of "classroom" tests. These individuals have direct and primary responsibility for the content, quality, and efficacy of tests used in a particular course. Such individuals carry such titles as course director, course coordinator, or senior instructor or they have been delegated the responsibility by a person carrying such a title. While titles vary, it is usually clear, to those directly concerned, who has operational responsibility for the testing done in a course. These individuals should become intimately familiar with the entire contents of this paper.

I should like to express my thanks to the chairman and members of the U.A.M.S. College of Medicine, Council for Academic Affairs, Committee on the Assessment of Student Performance; including, John Barnett, C. Bhuvaneswaran, D.D. Christ, Ross Dykman, Roberta Monson, Marvin Murphy, Mary Ragsdill, and Ture Schoultz for their careful reviews and constructive criticisms of earlier drafts of this paper. Also, a special note of thanks to Carolyn Cason, U.A.M.S. College of Nursing, for her technical review of this paper.

G.J.C.
Little Rock
4-1978

TABLE OF CONTENTS

PREFACE i

** I. INTRODUCTION 1

*** II. CONSTRUCTING THE TEST 3
*** A. General Considerations 3
*** 1. Consistency 3
*** 2. Cognitive Levels 3
* B. Test Construction Procedure 4
* 1. Test Plan Table 4
* a. Row Labels: Test Outline 4
* b. Column Headings: Cognitive Levels 4
* (i) Knowledge/Memory 5
* (ii) Comprehension/Understanding 5
* (iii) Application of a Specified Rule 5
* (iv) Problem Solving 5
* c. Number of Items 5
* (i) Per Cell 5
* (ii) Per Row 5
* (iii) Accuracy 5
* 2. Item Types 6
*** a. Response Modes 6
* (i) Constructed Response Items 6
* (ii) Selected Response Items 6
*** (iii) Response Mode and Cognitive Level 7
* b. Item Forms 7
* (i) Completion Items 7
* (ii) Fill-in-the-blank Items 8
* (iii) True-False Items 8
* (iv) Matching Items 8
* c. Multiple Choice Items 9
* (i) Simplicity of Expression 9
* (ii) Correct Grammar 9
* (iii) Universals and Absolutes 9
* (iv) Length of Options 10
* (v) Repeated Words and Phrases 10
* (vi) Number of Options 10
* (vii) Simple versus Complex M-C 10
* (viii) Incorrect Options 11
* (ix) Number and Position of Correct Options 11
* (x) Multiple-Possible-Correct M-C Items 12
*** 3. Item Writing 12
*** 4. Item Review and Selection 14
*** a. Review and Selection Responsibility 14
*** b. Criteria 14
*** (i) Ambiguous Items and Options 14
*** (ii) Consistency of Items with Instruction 15
*** 5. Final Copy: Preparation and Reproduction 15
a. Grouping Items 15
b. Response Instructions 15

* Information most relevant to test item writers.
** Information most relevant to educational administrators.
*** Information relevant to both of the above.

* c. Balance Correct Response(s).....15
d. Format: Items and Instructions.....15
e. Cover Sheet.....15
f. Proof Reading.....16
g. Reproduction of Test.....16
h. Answer Key.....16

** III. INFORMING STUDENTS BEFORE THE TEST.....24
** A. Formal Guidelines.....24
** B. Test Plan.....24
** C. Example Items.....24
** D. Testing Conditions.....24
** 1. When and Where.....24
** 2. Duration.....25
** 3. Required Materials.....25
** 4. Required Information.....25
** 5. Permitted Resources.....25
** 6. Prohibitions.....25
** E. Scoring Procedures.....25
** 1. Machine/Manual.....25
** 2. Faulty Items.....25
** 3. Computation of Test Score.....25
** 4. Combining Scores Across Tests.....26
*** F. Evaluation/Grading Procedures.....26
*** G. After Test Feedback.....26
** 1. Student Verification.....26
** 2. After Test Discussion Session.....26
** 3. Test and Item Analysis Information.....26

IV. GIVING THE TEST.....27
** A. Conditions of Testing.....27
** B. One-Shot Tests.....27
** C. Repeated Tests.....27
1. Introducing Repeated Tests into a Course.....28
a. Introduce Repeated Unit Tests.....29
b. Expand the Number of Alternate Test Forms.....29
c. Convert Review Quizzes to Repeated Tests.....30
d. Wholely Repeated-Tests Based Course.....31
2. Writing Equivalent, Alternate Test Forms.....31
3. Test Security.....32
4. Establishing a Criterion Score.....32
5. Scheduling Repeated Tests.....34
6. Administration of Repeated Tests.....34

** V. SCORING AND ANALYZING THE TEST.....36
A. Item Difficulty.....36
1. Calculating the Difficulty Index.....37
a. Right-Wrong Scoring.....37
b. Partial Credit Scoring.....38
2. Interpreting the Difficulty Index.....39
a. Critical Material Items.....39
b. Representative Material Items.....40

* Information most relevant to test item writers.
** Information most relevant to educational administrators.
*** Information relevant to both of the above.

B. Item Consistency....43
1. Calculating the Discrimination Index....44
2. Interpreting the Discrimination Index....45
a. Critical Material Items....47
b. Representative Material Items....47
*** C. Scoring Faulty Items....47
* 1. Invalid (No Credit) Items....48
* 2. Bonus Credit Items....48
* 3. Combination of Invalid and Bonus Items....49
* 4. Partial Credit Items....49
5. Limitations on Deletions of Faulty Items....49
** D. Test Reliability....50
** 1. Appropriate Reliability Coefficient Values....50
** 2. Reliability versus Validity....51
3. Finding the Approximate Reliability of a Test....52
a. Estimating a Test's Standard Deviation....52
b. Method 1 for Approximating the Test's Reliability....52
c. Method 2 for Approximating the Test's Reliability....53

** VI. INFORMING STUDENTS AFTER THE TEST....55
** A. Information to be Provided....55
*** B. After Test Discussion Session....55
** C. Confidentiality....55

BRIEF BIBLIOGRAPHY....56

* Information most relevant to test item writers.
** Information most relevant to educational adminstrators.
*** Information relevant to both of above.

GUIDELINES FOR OBJECTIVE TESTING OF STUDENTS

Gerald J. Cason, Ph.D.

Educational Development, UAMS

March 1978

I. INTRODUCTION

The following guidelines were developed specifically for cases in which objectively scored written tests are to be used.

Because the great majority of tests built and given by instructors are intended to measure students' cognitive performance, the following focuses upon cognitive testing. Cognitive performance includes all those intellectual activities which primarily involve the manipulation or processing of information. Commonly cited levels of cognitive performance include (a) Memory/Knowledge i.e., the recall of information; (b) Comprehension/Understanding of information; (c) Application of rules and principles; and (d) Problem Solving which includes selecting the appropriate rule(s) in addition to Memory, Comprehension, and Application.

Tests of cognitive performance are usually composed of a collection of items, problems, or questions to which the student can make a relatively simple, unambiguous response such as (a) marking a box on an answer sheet; (b) circling a letter; (c) writing a single letter, word, or short phrase, or (d) writing the numbers and symbols at each step in finding the solution to a computational problem. Written items requiring such unambiguous "written" responses are used for reasons of speed, ease, and most important, accuracy in scoring. Such tests are commonly called "objective tests" largely because the rules for scoring them are simple enough that different people will arrive at the same score for a given set of student responses and the same "correct answer key". Indeed, the scoring rules for an objective test are so simple that usually it is very little trouble to use a machine to score it. It should be clearly understood that the objectivity of an objective test usually derives more from the process of scoring than from the manner in which items are written or selected.

When objective cognitive tests are properly built and used, they can serve four educationally important functions. These tests provide the instructor information on what a student can and cannot do. The instructor needs this information both to (a) select and/or design further instruction and (b) decide if and/or when a student is qualified for advancement to the next unit, block, course, or level of instruction. Well constructed and properly used objective cognitive tests provide the student (a) a guide to

study and (b) confirmation of what he is able to do. The following guidelines were developed to assist instructors to construct and use objective cognitive tests to most effectively serve all of the four functions mentioned above. When an instructor builds or uses an objective test, he should make every reasonable effort to comply with all the guidelines provided below.

The following sections provide guidelines for (a) constructing the test; (b) informing the students about the test; (c) giving the test; (d) scoring and analyzing the test; and, (e) reporting results of the test. In addition to the guidelines, some explanatory material and supporting rationale is provided.

II. CONSTRUCTING THE TEST

A. General Considerations

1. Consistency.

A test should be clearly consistent in both content and emphasis with (a) the content and emphasis of the instruction actually received by the students, and, (b) the content and emphasis that the students were informed the test would contain.

Content means the topics or instructional objectives covered as well as the cognitive level of the information. Emphasis in the test refers to the number of points associated with different topics or cognitive levels; while in instruction it refers to time or stress given. In this context, instruction means lectures, seminars, conferences, exercises, and assignments.

If instruction were primarily concerned with "learning facts", then the test should primarily measure student knowledge or recall of those "facts" emphasized during instruction. Conversely, if students received little or no instruction in how to select among several rules/principles to solve a problem, there should be respectively few or no items in the test which requires this type of cognitive activity.

2. Cognitive Levels.

Cognitive processes include (a) Memory/Knowledge, (b) Comprehension/Understanding,(c) Application, and (d) Problem Solving. These form a progression of increasing complexity with each higher level involving some reliance upon performance at all lower levels. Because of this hierarchical relationship, it is possible to indirectly measure performance on all cognitive levels by a test composed solely of Problem Solving items. More generally, if you construct a test solely of items at the highest cognitive level in which you are interested (e.g. Application), then the items on this test will also measure student performance on relevant information at all lower levels (e.g., Memory/Knowledge and Comprehension). However, as a practical matter, it may be difficult to devise test items at one level that adequately cover (i.e., sample) the breadth of material at lower levels in which you are interested. For this reason it is usually necessary to include some items at each level below the highest cognitive level in which you are interested.

While items at a higher level provide some information about performance at lower cognitive levels, as you progress from a lower to a higher level, something new is

added. Thus, adequate performance on a test composed of items measuring Memory/Knowledge, Comprehension, and Application does not necessarily indicate that a student would perform well on Problem Solving items (i.e., the next higher level).

Taken together, the above two points mean that in virtually all cases you will need to include items which measure performance at the highest cognitive level in which you are interested and items at each lower level. Adequate performance at all lower levels is not sufficient to demonstrate adequate performance at the higher level. Conversely, adequate performance at the higher level is not usually sufficient to demonstrate adequate achievement on lower levels across all the material (i.e., information) which you may wish the test to cover.

B. Test Construction Procedure

An instructor should use the following procedure (or one that has substantially the same results) to design and produce an objective test of cognitive performance.

1. Test Plan Table.

Construct a table in which each row corresponds to a single topic (subtopic) or instructional objective and each column corresponds to a different level of cognitive complexity.

a. Row Labels: Test Outline. Write an outline of the content to be covered by the test using either (a) one topic/subtopic per row; or (b) one instructional objective/subobjective per row. The outline should divide the content into sufficiently small pieces that you will need to use only a few (2-7) questions for each row and yet still obtain an adequate coverage of the material within each topic (sub-topic) or objective (sub-objective).

b. Column Headings: Cognitive Levels. Write column headings for all the cognitive levels for which items will be included in the test. Arrange the cognitive level names in ascending order of complexity from left to right. See the example test plan provided in Table 1.* For example, you may wish to use the following level names and descriptions (cf. Bloom, 1956):

* Table 1 and example items may be found at the end of this section i.e., beginning on page 17.

(i) Knowledge/Memory: items requiring student to recall information (facts, principles, relationships) in order to either construct a correct answer or select the correct answer from a given list. See Examples 1 and 2.*

(ii) Comprehension/Understanding: items requiring the student to demonstrate that he can either construct or select a statement which has a meaning equivalent to that of a given statement. See Examples 3 and 4.*

(iii) Application of a Specified Rule or Principle: items requiring the student to use a specific designated rule, principle or procedure to solve a problem or answer a question. See Example 5.* The student may be required either to construct the "answer" or select from a list of alternatives.

(iv) Problem Solving: items requiring the student to recall a whole set of rules or principles, choose the appropriate ones to use, and apply them in the correct order to solve the given problem. See Example 6.*

c. Number of Items. Assign a number from zero to seven to each cell in the Test Plan Table. The number assigned to each cell (i.e., intersection of a row and column) specifies how many items will be included in the test which are at the cognitive level defined by the column heading and on the topic (subtopic) or objective (sub-objective) defined by the row label.

(i) Per Cell. The number assigned to each cell should be proportional to the importance of the corresponding column heading and row label.

(ii) Per Row. If you have a total of more than seven items on a single row, you should make finer divisions of topics/objectives in the outline (row labels).

(iii) Accuracy. The larger the number of items used to measure something, the more accurately it will be measured. Thus, in general you should use as many items as is practical (i.e., the longer the test the better).

* Table 1 and example items may be found at the end of this section i.e., beginning on page 17.

2. Item Types.

a. Response Modes. If the test is to be objectively scored, the items must be such that they require responses from the student which are relatively simple and unambiguous. That is, essentially no "judgement" should be needed to determine if a student's response is correct or incorrect. Within this general requirement two response modes are usually acceptable, (a) Constructed Response Items and (b) Selected Response Items.

(i) Constructed Response Items. This type of item requires the student to construct, write or draw a short unambiguous answer such as a single word or phrase, simple diagram, or a series of numbers (e.g., in a computation problem). See Examples 1 and 3. Care must be taken to avoid writing items to which ambiguous answers may be provided by the student. If this occurs, the objectivity in scoring can be greatly reduced or entirely destroyed. The advantage of this type of item is that relatively good items can usually be written fairly easily and quickly. However, scoring tends to be quite time consuming if a large number of items and/or students is involved. As a rule of thumb, if the total number of test items times the total number of students to be tested is less than about 200 then Constructed Response Items may be used.

(ii) Selected Response Items. This type of item requires that a student select the correct, best, or most appropriate response to an item (stem) from among a given list of responses (options). Good items of this type are relatively more difficult to write than Constructed Response Items. However, because such items can be easily presented as some variation of multiple-choice or matching format, Selected Response Items are easily machine scored. Machine scoring not only greatly increases the speed and ease of scoring, usually the computer program used for machine scoring provides test analysis information at essentially no additional cost. (This is the case with the machine test scoring service provided by the UAMS Computing Facility, i.e., OPSCAN). Whenever the number of items in the test times the number of students who are to take it exceeds about 200, you should use Selected

Response Items (regardless of whether the test will be hand or machine scored).

(iii) Response Mode and Cognitive Level. Although there is a widely held notion that Constructed Response Items are better (than Selected Response Items) for measuring performance at the higher cognitive levels, the extensive empirical research on this issue has not provided any convincing support for this notion. On the contrary, the research indicates each of these two response modes are equally appropriate for all cognitive levels. If you were to construct two equivalent tests, each test using only one of the two response modes, the results of the two tests would be highly correlated. This means that students who score high on one item type strongly tend to score high on the other and similarily those who score low on one score low on the other. Furthermore, there is no difference in the information provided by these two item types with respect to long term recall (retention), transfer of learning, or any of a dozen other aspects of learning which have been investigated. Thus, the choice should be made on the basis of relative effort needed to write, score and analyze the number of items needed for a test of a group of students of a given size.

b. Item Forms. There is an indefinitely large number of ways in which objective items can be arranged (whether requiring constructed or selected responses). The most common are completion, fill-in-the-blank, true-false and matching. In general, you should use whatever form(s) that (a) make it easiest for you to write items which test what you want them to test, (b) reduce the probability of a student's guessing correctly and (c) avoid complex instructions.

(i) Completion Items. These items require the student to construct or select information which correctly completes the (incomplete) statement provided in the item stem. Alternatively, the stem may be phrased as a question and the student must construct or select the correct answer(s). See Examples 1, 2, 3, 4 & 6.

(ii) Fill-in-the-blank Items. These items usually provide a statement with some critical word(s) or phrase(s) omitted. The student's task is to insert the missing elements either by a constructed or selected response. See Example 5. Unless care is exercised in using this arrangement you may produce items which either are trivially easy or which are very ambiguous. The first difficulty is easily avoided as it usually arises from the omission of non-critical words or phrases. The second arises most often from the deletion of too much information, (e.g., The _____ states that _____will usually_____ when the _____ is _____).

(iii) True-False Items. These items involve nothing more than a stem containing a complete declaratory statement which the student must classify as either true or false. The major shortcoming of this item type is that the student has about a 50% probability of getting the correct answer simply by guessing; therefore, a large number of T-F items are required to provide useful information. Also, in creating false statements, it is sometimes difficult to avoid producing items which students view as "trick" questions. When this type is used make sure you have about an equal number of true and false statements. Care should be exercised in avoiding the use of negatives (particularly multiple negatives) in T-F items.

(iv) Matching Items. Depending upon how generally you want to construe the term "matching" you might wish to include all forms of Selected Response Items. However, usually "matching" is used to denote items in which the student is presented with two or more lists of things and he must correctly match the things in one list with the things in the other(s). Generally, you should make one list longer than the other to prevent the student's selecting the last match by the process of elimination. See Example 7. This applies even in the case where you wish the student to place the steps of a procedure in the correct order. In this case, you might include some things that were not actually steps in the designated procedure.

c. Multiple Choice (M-C) Items. This item type is most appropriate when you have large numbers of items and students, i.e., when the Selected Response Mode is appropriate. All or most of your items should be M-C whenever the number of students to be tested times the number of items to be used substantially exceeds 200. M-C items are particularly appropriate in this case because (a) they are compatible with machine scoring and analysis and (b) all of the item types described above (as well as others) can be presented in M-C form. The principal limitation of M-C items is the limitation imposed on simple matching when machine scoring is used. Usually you can have only five options for each item on a machine scored test. (This is the case at UAMS). This may cause some difficulty by restricting the length of one list to five things. See Example 7. (The UAMS machine scoring procedure provides a way around this but some users have found it unsatisfactory). If you use M-C items (whether machine scored or not), the following guidelines should be strictly observed:

(i) Simplicity of Expression. Use as simple grammar and vocabulary as is possible in both the item stem (question) and response options (answers). Avoid the use of multiple negatives and other needlessly complex grammatical constructions.

Poor: It is not infrequently the case that desert flowers bloom after a rain shower.

Better: It is frequently the case that desert flowers bloom after a rain shower.

Better Yet: Desert flowers frequently bloom after a rain shower.

(ii) Correct Grammar. Write item stem (question) and options (answers) so that the combination of the stem is grammatically correct with each option.

(iii) Universals and Absolutes. Except where actually required, avoid the use of universals and absolutes, e.g., all, always, never, perfect, exactly, etc.

(iv) Length of Options. Write options such that they are all approximately the same length.

(v) Repeated Words and Phrases. Avoid writing options in which some contain the same word(s) or phrase(s) while others do not. Frequently when this is done the instructor has inadvertently supplied information about which sub-set of the options contains the correct answer. See Example 8. If a word or phrase is contained in all the options and is either at the beginning or end of every option, it should be placed in the item stem. See Examples 9 and 10.

(vi) Number of Options. Just because the standard answer sheet provides for five options do not feel obligated to always use five. Four good options are better than five options where one will almost never be choosen by a student. The objectives served by having multiple options (choices) are to (a) reduce the probability that a student can choose the correct answer by simply guessing and (b) require the student to make meaningful discriminations between the given options. If virtually all students can recognize one (or more options) as being obviously incorrect, little or nothing is gained by its (or their) inclusion. The phrases "all of the above" and "none of the above" and similar phrases should be strictly avoided as options. If you really want to know if the student recognized each option as correct or incorrect use the multiple-possible-correct M-C item type. Never use these phrases simply to fill out the number of options.

(vii) Simple Versus Complex M-C. Use the simplest form of M-C that meets your needs. Usually, "single correct or best answer" or "multiple possible correct" M-C items should be used (Examples 6 and 11) rather than more complicated forms (Example 12). It is much more difficult to write good quality items of the more complex forms of M-C. M-C multiple-possible-correct items (e.g., Example 11) are thought to be more fair by students. Also, M-C multiple-possible-correct items provide a substantial advan-

tage in adjusting student scores for "faulty" items, e.g., when a student can provide a convincing argument that one of the options intended as incorrect is as good or better than the option(s) intended to be correct. However, this has to be balanced against the difficulty sometimes associated with balancing the number of correct options found across a set of "multiple-possible-correct" items. See section (x) below. The complex-matching M-C items such as Example 12 should be avoided because of the complexity of the response coding and instructions necessary to answer such questions. The more complex the response instructions become the more likely students are to make errors which are more related to their compliance with test instructions than true competence in the area the test is intended to measure.

(viii) Incorrect Options. The incorrect option(s) should be selected (by the instructor) for some reason beyond simply that they are wrong. Preferably, separating the correct option(s) from incorrect should require that the student make non-trivial discriminations. Incorrect options should represent some kind(s) of error (i.e., misunderstanding, misinformation) that a student might conceivably make. For example, in an item requiring the student to select the most appropriate therapy for a given case, some of the (incorrect) options might be correct in the "general case" but wrong in the given instance because of secondary complications or, the patient's age, weight, drug interactions, etc.

(ix) Number and Position of Correct Option(s). There should be no pattern in which particular option(s) is (are) correct. That is, which option(s) is (are) correct should be random with respect to its (their) location in the list of options. (a) Single correct or best answer M-C items: Across all items of this type in the test, each option (A, B, C, D, E) should be correct about an equal number of times. (b) True-False (T-F) M-C Items: When A=True and B=False, there should be approximately the same number of items for which A and B are correct. That is, approximately an equal number of true and false items.

(x) Mutiple-Possible Correct M-C Items. There should be about the same number of items in which each option is correct. That is, the frequency with which "A" is correct should be about equal to the frequency with which "B", "C" etc. are each correct. Furthermore, the number of items in which no option is correct should be about equal to the number of items in which 1 option, 2 options, etc. are correct. These requirements are the primary reason that using this type of item is sometimes more difficult than the "single best answer" type. However, the whole point of balancing the number and position of "correct" options is to avoid providing unintentional clues to examinees. For best effect, multiple-possible-correct items should be presented so that the student thinks that for an item selected at random there is about an equal probability that none, one, two, three, four, or all options are correct. Furthermore, the student should think that the position of correct option(s) is random. If there is a pattern to how many or which option(s) is correct, students will use this information to improve their chance of guessing correctly.

3. Item Writing.

Write a sufficient quantity and kind of items to satisfy the requirements of your test plan. As you write items try to insure that they conform to the guidelines provided above. Where possible, several more items per cell than are called for in the test plan should be prepared. This will permit selection of better quality items. Who actually writes the items is less important than that the items conform to the guidelines. A deliberate effort should be made to prepare items of intermediate difficulty. The intuitively attractive notion that there should be some easy, some moderately difficult, and some hard items on a test is not supported by experimental evidence. On the contrary, when the purpose of the test is to detect differences between students taking it, the greatest spread of scores is obtained when only items of moderate difficulty are used. See Figure 1.

You should write each item on a separate card. This will facilitate the final review, selection, and organization

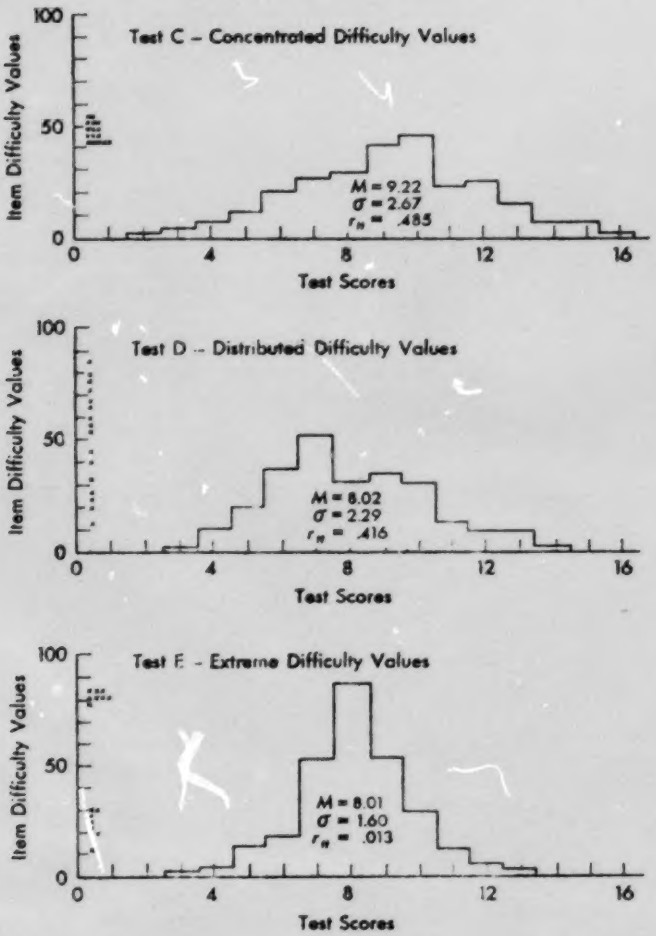


Figure 1. Relation between Distribution of Test Scores and Distribution of Item Difficulty Values. (From Ebel, 1965, p. 363.) Small X's at extreme left in each graph indicate difficulty of individual items in each test.

M = Mean (average) test score
σ = Standard Deviation of test scores
r_{tt} = Reliability Coefficient of test

ED/3-78

process. Also it will permit you to build a file of items for use on future tests. Test analysis data such as the item's Difficulty Index and Discrimination Index, number of students in the test group, description of the students in the test group (e.g., Sr. Med. Students), and the date of the test should be recorded on this card each time the item is used on a test.

4. Item Review and Selection.

The following process should be used to obtain items for use both in the test and as examples during instruction:

a. Review and Selection Responsibility. Final authority and responsibility for the review and selection of items should always rest with a member of the faculty who was present during the instruction covered by an item. While this will frequently mean the instructor who actually conducted the relevant instruction, it could be a faculty member who was present for other reasons. This requirement is needed to assure that an item is indeed consistent with what was taught. This, of course, does not mean that an instructor should be free to include instruction on material and test items on material which falls outside of any established curricular guidelines on the appropriate content of the course for which the test is intended.

b. Criteria. Select items which best conform with (a) the test plan specifications, (b) item guidelines above, and, (c) curricular guidelines which apply to the course; and, if available from previous use, (d) have satisfactory Difficulty and Discrimination Indices. (See sections in II above and V below.)

(i) Ambiguous Items and Options. Be especially careful to check items for ambiguities. Discard or revise any item which would permit an option intended to be wrong being arguably correct. Where possible at least one person other than the item's author should review it. This might be either a faculty colleague or an advanced student.

For constructed response items, it is necessary to write the correct answer key prior to the item review and selection process. This "key" should contain "model" answers to test items. These model answers are used as standards against which the student answers will be evaluated.

(ii) Consistency of Items with Instruction. Strive to insure that the intended correct answer(s) corresponds to what the actual instruction would lead a student to believe is correct. If reviewers cannot agree among themselves, the item should be revised until they do or it should be discarded. If there is a conflict among several sources of information, the item should make clear which source is to be used to answer. The item review should embrace all the guidelines provided above as well as possible grammatical, typographic and substantive errors in items.

5. Final Copy: Preparation and Reproduction

a. Group Items. Group items of the same type together e.g., place all "single-correct M-C" items together, all T-F items together, etc.

b. Response Instructions. At the beginning of each section of the test containing a single type of item, insert instructions for answering these items. If instructions are particularly complex, e.g., instructions for complex-matching M-C items (see Example 12), these instructions should be repeated on each page containing this item type.

c. Balance Correct Response(s). Re-arrange items and/or options within items so pattern of correct responses is random. See sections (ix) and (x) above.

d. Format: Items and Instructions. Use a uniform method for placing instructions, item stems, and options on page. When machine scoring is used, be sure that response options for an item are identified by a letter (A, B, C, D, E) if letters are used on the answer sheet or a number (1, 2, 3, 4, 5) if a number is used on the answer sheet. Do not "split" an item across the end of a page, i.e., do not put part of the item at the bottom of one page and the rest of the item on the following page.

e. Cover Sheet. Prepare a cover sheet which provides the following information:

(i) Number and type of items included;

(ii) General instructions on how student is to record his answers;

(iii) Any particular information student is to put on answer sheet, e.g., name, identification number, where it goes and how it should be entered;

(iv) Time limit (where applicable); and,

(v) Any other general information needed by students to take test properly.

f. Proof Reading. Have draft copy of the test written or typed in accord with all of the above guidelines. Carefully proof-read draft and correct any errors (clerical or otherwise). Each time you have some portion of a test re-typed, be sure to re-proof that portion prior to reproducing the test. Although a good secretary or clerk typist may seldom make an error, the instructor has the final responsibility for the accuracy of a test (in the same way the author has final responsibility for the content of an article; not the journal editor, nor the secretary that typed it, nor colleagues who reviewed it.)

g. Reproduction of Test. Reproduce the test in sufficient quantities for your needs. Generally, care should be observed to prevent the specific contents of the test from being available to those who are to take it. Common sense precautions are usually sufficient e.g., do not leave copies of the test lying about where students may have access to them.

h. Answer Key. Always prepare a correct answer key prior to giving the test.

Table 1

Schematic Test Plan Table

Instructional Objectives or Topics	Level of Cognitive Complexity				
	1 Knowledge/ Memory	2 Comprehension/ Understanding	3 Application of Single Rule	4 Problem Solving	Row Totals
I. (15)					
A.	2	3			5
B.	4	1			5
C.	1		2	2	5
II. (10)					
A.	1		2	1	4
B.	1		4	1	6
III. (23)					
A.	1	2	1	3	7
B.					
1			2	5	7
2			2	4	6
3.	1			4	5
Column Totals (50)	11	6	13	20	50

Note: In the rightmost column are entered the names or descriptions of the instructional objectives or topics covered by the test. Cell entries specify the number of items the test will contain for each combination of Instructional Objective and Level of Cognitive Complexity. The distribution of items should reflect the relative importance of different topics and levels of cognitive complexity. In this example topic III and Level 4 (i.e., Problem Solving) are most important.

EXAMPLE 1. Knowledge/Memory Item: Constructed Response.

STEM: Briefly state Newton's "Third Law" as given in your text.

RESPONSES: To be supplied (constructed) by student.

EXAMPLE 2. Knowledge/Memory Item: Selected Response.

STEM: Which of the following correctly states Newton's "Third Law"?

RESPONSES:

A. A body at rest tends to remain at rest until acted upon by some (outside) force.
B. For every action there is an equal and opposite reaction.
C. Thermal processes tend toward the equal distribution of kinetic energy.
D. In a series of unfolded sedimentary beds, the oldest material is the lowest in the series.
E. For any object whose density is less than water, its weight equals the weight of the water it will displace.

NOTE: In both Examples 1 and 2, the student must recall the statement of Newton's "Third Law"; however, neither example requires that he be able to either understand or apply it. In Example 1, the student must be able to construct (write) the statement of the Third Law as given in his text. In Example 2, he must be able to select the correct statement from a given list of alternatives (options).

EXAMPLE 3. Comprehension/Understanding Item: Constructed Response:

STEM: In your own words, state the Law of Superposition.

RESPONSES: To be supplied (constructed) by student.

EXAMPLE 4. Comprehension/Understanding Item: Selected Response.

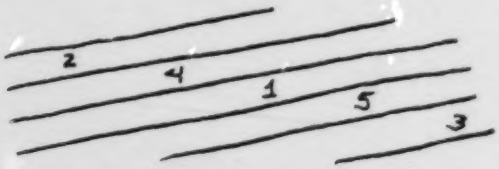
STEM: The Law of Superposition means about the same as:

RESPONSES:
A. Any explanation is better than none.
B. Deeper rock is usually older.
C. Rank has its priviledges.
D. Water seeks its own level.
E. Still water runs deep.

NOTE: Both Examples 3 and 4 require that the student both recall the Law of Superpositon and either construct or recognize a statement with equivalent meaning.

EXAMPLE 5. Application of a Specified Rule: Single Best Answer.

FIGURE:



STEM: In the above crossectional illustration of sedimentary beds, according to the Law of Superposition, bed _____is probably the oldest.

RESPONSES:
A. 1
B. 2
C. 3
D. 4
E. 5

NOTE: In Example 5, the stem provides the name of the specific principle to be used in answering the question. The student must be able to recall the statement of the designated principle and understand its meaning in order to apply it to the given situation.

EXAMPLE 6. Problem Solving: Single Best Answer.

STEM: While you are cross-country skiing, you come upon the scene of an auto accident in which four people have suffered injuries. You have no medical equipment or supplies and the nearest assistance is at least six (6) hours away. Given the following description of each person's condition (and assuming no other problems are present) which person should you help first

RESPONSES:
A. 42 year old male exhibiting all clinic symptoms of cardiac arrest but not get ting noticeably cyanotic.
B. 37 year old pregnant female in last moments of normal labor.
C. 10 year old male with compound fracture of femor and "spurting" blood from laceration of neck.
D. 8 year old female with 2nd degree burn on less than 10% of her body but whose clothing is still aflame.

NOTE: In Example 6, no indication is given of what rule or rules should be applied, nor the sequence in which the rules should be used. In addition, the kinds of rules which might be applicable range from those of a technical/professiona nature to those of a moral/ethical nature. In order to solve this sort of pro lem the student must remember the relevant facts, principles and procedures, understand them well enough to select the appropriate ones and apply them to a rather complex case which may be substantially different from any case he has previously encountered.

EXAMPLE 7. Simple Matching.

DIRECTIONS: Which one of the options below (A-E) is most closely related to the development or modern use and limitations of each of the things named/described in items 26-30. (Each option may be used more than once).

RESPONSES:
A. Principle of Superposition
B. Newton's Third Law
C. Ohm's Law
D. Pythagorean Theorem
E. Goedel's Theorem

STEM(S):
26. Celestial navigation
27. Cryptography
28. Rocket motor
29. Electric motor

EXAMPLE 8. Repeated Words and Phrases.

DIRECTIONS: Each of the questions or incomplete statements below is followed by five suggested answers or completions. Select the one that is BEST in each case.

STEM: A molar solution of an electrolyte has a higher osmotic pressure than a molar solution of a non-electrolyte because the _____.

RESPONSES:
A. electrolyte ions carry electric charges.
B. electrolyte ions are usually hydrated.
C. electrolyte alters the dielectric constant of water.
D. electrolyte dissociates in solution.
E. non-electrolyte is non-polar.

NOTE: In Example 8, some words are repeated in several but not all options. Many students are aware that when this occurs the correct answer is frequently one of the options containing the repeated words or phrases.

EXAMPLE 9. (POOR) Repeated Word.

STEM: In basal fractures of the skull, if cerebrospinal fluid escapes from the nose, there is probably a fracture of the _____.

RESPONSES:
A. nasal bone
B. temporal bone
C. ethmoid bone
D. parietal bone
E. frontal bone

EXAMPLE 10. (BETTER) No Repeated Word

STEM: In basal fractures of the skull, if cerebrospinal fluid escapes from the nose, there is probably a fracture of the _____bone.

RESPONSES:
A. nasal
B. temporal
C. ethmoid
D. parietal
E. frontal

NOTE: The word "bone" has been repeated at the end of all the options in Example 9. In Example 10, the word "bone" has been moved into the item stem.

EXAMPLE 11. Multiple-Possible-Correct.

DIRECTIONS: In the following item none, one or more options may be correct. Mark all correct options.

STEM: Which of the following cities is a (are) state capital(s)?

RESPONSES:
A. Little Rock, Ark.
B. Los Angeles, Calif.
C. Austin, Tex.
D. Albuquerque, N.M.
E. Oklahoma City, Okla.

EXAMPLE 12. Complex Matching.

DIRECTIONS: In the item below one or more options below is correct, mark:

RESPONSES:
A. if only 1, 2 and 3 are correct.
B. if only 1 and 3 are correct.
C. if only 2 and 4 are correct.
D. if only 4 is correct.
E. if all four (1,2,3, & 4) are correct.

STEM: Which of the following elements is (are) normally present in sea water in quantities exceeding 100 ppm?

OPTIONS:
1. Magnesium
2. Bromine
3. Sodium
4. Sulfur

NOTE: If the student can eliminate only "2" as incorrect, he need only guess between C and D. If he can only eliminate "4" as incorrect, he must guess from A and B.

III. INFORMING STUDENTS BEFORE THE TEST.

Students should be provided all the following kinds of information. This information should be provided in writing. It should be provided early enough and in sufficient detail for students to use it as a realistic and accurate guide for study.

A. Formal Guidelines.

If some formal testing guidelines are used (e.g., such as the guidelines in this document), then students should be informed of this and provided access to the document(s) containing such guidelines.

B. Test Plan.

Students should be provided a copy of the test plan(s) which will be used to construct any test they will take. The test plan they receive should contain the same information as the one used by the instructor: topics/objectives, cognitive level of items, point value and number of items in each cell of table. (See section II. B. 1. above.)

C. Example Items.

Students should be provided an example of each item type that will be included in the test. For each cell in the test plan from which one or more items will be drawn for the test, at least one example item should be provided. Correct answers should be given with the examples. An explanation as to why various options are correct or incorrect should be provided if students request this information. Problems from exercises or old tests may be used for this purpose. Students should be told which example is intended to exemplify each cell in the test plan from which items in the test will be taken. As far as possible, example items should be of moderate difficulty, i.e., neither unusually easy nor difficult. An item is of moderate difficulty when 30-70% of the students would answer it correctly if it were included in the first test they attempted on the material (and had not been used specifically as an example).

D. Testing Conditions.

Students should be informed of the conditions under which the test will be given. This should include the following items of information:

1. When and Where.

At least the approximate date, time, and location (e.g., room number) of the test(s) session.

2. Duration.

The length of time allowed to complete the test.

3. Required Materials.

What materials/equipment the student must provide, e.g., pencil, pen, typewriter, paper, microscope, pocket calulator, etc.

4. Required Information.

Specific information the student must bring with him/her e.g., student number.

5. Permitted Resources

What materials/equipment/information the student will be permitted to use such as reference books, lecture notes, checklists, calculators, etc.

6. Prohibitions.

What actions/activities will be prohibited during the test, e.g., using notes, confering with other students, leaving test room.

E. Scoring Procedures.

Students should be informed of the explicit scoring procedures that will be used. This should include the following information:

1. Machine/Manual.

Whether machine or manual scoring will be used and any special requirements or conditions this may impose. For example, stray marks on machine scored answer sheets may cause the loss of points.

2. Faulty Items.

How questionable or faulty items will be identified (e.g., through student challenge of the "correct answer" key and/or use of item discrimation index). How the identified faulty items will effect student scores (e.g., all students given credit or item disregarded).

3. Computation of Test Score.

How student test scores will be computed from scores on individual items, e.g., arithmetic sum of points, Z-scores or other standard scores.

4. Combining Scores Across Tests.

What procedure will be used to combine the score on this test with scores on other tests including any weighting scheme to be used.

F. Evaluation/Grading Procedures

Students should be told the explicit rules or procedures and standards which will be used to convert the score on this test alone (or in combination with other test scores) into grades such as PASS/ FAIL and/or A, B, C, D, etc.

G. After Test Feedback

Students should be told when and how they will receive information regarding their performance on the test.

1. Student Verification

Provision should always be made for students to go over their own test, item by item, to see if the "correct answer" key was properly used to score their test. This applies whether manual or machine scoring is used. Students should also be told if they will have an opportunity to challenge the answer key.

2. After Test Discussion Session.

An item by item after-test discussion session should be conducted as it will be useful to both the instructor and students. Students need to be told why incorrect answers they gave were incorrect. Emphasis during this discussion should be placed upon (a) items which a large percentage of students missed, (b) items which had negative discrimination indexes, and (c) items for which some students wish to challenge the answer given in the key. Use of this procedure will tend to reduce conflicts between students and faculty regarding the fairness of tests if a satisfactory method for identifying and scoring faulty items has been adopted. Also, an after-test discussion session will help the instructor to determine why specific questions were missed. This will assist the instructor to improve and/or clarify parts of the instruction.

3. Test and Item Analysis Information.

If any test or item analysis information will be provided to students (e.g., distribution of student scores, item difficulties or discrimination indexes, reliability coeficient, etc.), this information should be provided either prior to or at the beginning of the after test discussion session.

IV. GIVING THE TEST

A. Conditions of Testing.

The test should be given under as nearly those conditions which students were informed it would as is practical. If practical circumstances require a major departure from the testing conditions which students were led to expect, the instructor should, if possible re-schedule the test.

None of the conditions under which the test is given should introduce irrelevant difficulty into the task. For example, if computational problems were involved and it were reasonable to expect that once "on the job" the student would be free to use a pocket calculator (or other job aide), the use of a pocket calculator (or other job aide) should usually not be denied during the test. More generally, unless you have specific need to know how well students perform under such adverse conditions as (a) excessive heat, cold, light, or noise; (b) inadequate light, time, space, equipment, or materials; (c) poorly written/printed materials; or (d) high physiological and/or psychological stress, then, reasonable steps should be taken to avoid these and similar conditions.

B. One-Shot Tests

In conventional practice, the test over a unit of instruction is given at one fixed time and place. Once taken and scored, the students score is "permanent and official". Except in rare instances no provision is made for the student to take another test (on the same material) to raise his score (or grade). Under these circumstances most students think the best strategy for improving their eventual course grade is to concentrate their efforts on studying the material that will be covered in later tests. The inclusion of a few review questions in later tests has little real effect on what the student studies. Most of the student's efforts are directed toward what is emphasized on later tests, i.e., material presented after that on which he made a low test score. For this reason "one-shot testing" fails to assure that a student will demonstrate some minimally acceptable level of achievement on each unit of instruction in a course. Thus, whenever possible, one-shot testing should be avoided in favor of the "repeated tests" approach.

C. Repeated Tests

The repeated tests approach initially requires more instructor effort than does the one-shot testing approach alone. However, repeated tests can produce much higher levels of quality assurance. Repeated tests, as the term suggests, require that a student be given multiple opportunities to take the test covering a unit, module, or block of instruction. This is accomplished in part by the instructor's constructing several equivalent, alternate forms of the test for a unit of instruction.

Greatest effectiveness is achieved from repeated testing when it is applied in conjunction with a minimum passing criterion score. That is, scores on repeated tests are interpreted as essentially Pass/Fail. In common practice the criterion is set high enough that only the more competent students pass the test. Operationally, students are required to take alternate forms of a test until they pass one of the alternate forms, i.e., until they achieve a score equal to or higher than the established criterion. Because the criterion score is usually set high enough that only those with fairly high levels of achievement can pass it, the repeated tests approach is sometimes called "mastery testing."

The effectiveness of repeated tests is also increased by designing them to cover smaller units of instruction than are commonly covered by hour quizzes in college courses. For example, a typical college course covering one semester and having three hours of credit associated with it will have three one-hour quizzes and a final exam. Each hour quiz will cover the material presented during about six weeks (i.e., eighteen lecture hours). The final exam covers all the material in the course (i.e., about 50 lecture hours). Repeated tests work more effectively when they cover no more than the material covered in 3 to 6 lecture hours (i.e., in one to two weeks in the typical college course). Thus, in the typical college course, the material should be divided into 9 to 18 units or modules with repeated alternate tests available for each unit. The test for an individual unit should, of course, be shorter than an hour quiz because it covers less material. Tests covering larger blocks of the course like hour quizzes should be continued. These block review tests should cover about 3 to 6 units. The more material in each unit, the fewer units should be covered by a review test.

1. Introducing Repeated Tests into a Course.

   Because writing good test items is such a time consuming process, generally repeated tests should be introduced into a course on an incremental basis. That is, the change over from one-shot testing to repeated testing should be accomplished in a planned evolutionary manner rather than trying to go from having one form of each test in the course one semester to many alternate forms of each test the next semester. Of course, the speed with which the change over may be accomplished is a function of many factors including the complexity of the subject matter, what kind of test item file exists (if any), and the resources available for the task (such as faculty time to write new items). For illustrative purposes the change over from one-shot to repeated tests

will be described in terms of a typical course which initially contains three one hour quizzes and a final exam, each of which is given as a one-shot test.

a. Introduce Repeated Unit Tests. For each block of instruction covered by an hour quiz, identify 2-6 coherent subunits of material. The first semester that repeated tests are used, write and give at least two alternate tests for each of the units covered by an hour quiz. Make full use of any item or test files that are available to provide items for the alternate forms of the unit tests. Continue to give the hour quizzes and the final exam as one-shot tests. For each unit, establish a test schedule indicating the first and last date upon which a student may take a test for that unit. For each unit, if a student fails to pass the first alternate test form, allow him to attempt to pass the second form of the test for that unit, unless the final date for taking tests on that unit has passed. Those students passing the first form are not permitted to take the second form of the unit test. Determine the student's final score or grade in the course on the basis of both the percent of unit tests he passed (irrespective of whether it was on the first or a subsequent attempt) and his scores on the one-shot "review" tests. In determining course scores or grades primary weight should be given to the one-shot tests, but enough weight should be given to the repeated unit tests that they are taken seriously by the students. There is little reason for repeated tests unless they provide a true incentive for the students to learn more. This incentive is provided to those who need it most (i.e., the students failing a unit test) by allowing them to study and retake a unit test to improve their course score. To accomplish this end, the percent of unit tests passed should count for at least 10% of the final course score or grade.

b. Expand the Number of Alternate Test Forms. On each subsequent occasion the course is given, following the first time repeated unit tests are used, one or more alternate tests should be added to the set for each unit. This process is continued until you have sufficient alternate forms of each unit test: usually between three and six. You have enough alternate forms when most students can retake tests until they have learned enough to pass the criterion score. There is a direct relationship between how much time you allow for students to take retests and how many alternate forms of the test will be required. Unfortunately, there are no general, empirically justified rules for determining how much time or how

many attempts should be allowed on each unit. This can only be determined by experience within a particular course. However, in general, the higher the level of competence you want to assure then the more time for re-study and the more alternate test forms for retesting must be used.

c. Convert Review Quizzes to Repeated Tests. After you have generated enough alternate forms of unit tests to meet your needs, you should next convert the "hour quizzes" (i.e., tests covering relatively large blocks containing several units) to repeated tests. In making this change, you should follow the same general approach as was specified above for introducing repeated unit tests. The first time around, provide at least two alternate forms of each review test. On subsequent occasions when the course is given at least one additional alternate form of each review test should be added until you have a sufficient number for your purposes.

A schedule specifying the first and last date upon which an alternate form of each review test may be taken should be developed. At this point an important change in how test results are interpreted may be introduced. If a student fails to achieve criterion on a review test by the last date on which he is permitted to take retests for that block, that means his progress in the course is unsatisfactory. If this happens on two successive review tests, then serious consideration should be given to having the student "drop" the course with a grade of Incomplete. In most courses, such a student is too far behind to catch up during the balance of the course. It would be much better for him to concentrate his efforts in other courses for the present and, at a later date, retake the course in which he is presently failing.

When the "hour quizzes" (i.e., review tests) are converted to the repeated test approach, final course scores should be determined on the basis of (a) percent of unit tests passed (regardless of number of attempts required), (b) percent of review tests passed (regardless of number of attempts required), and (c) the final exam which continues to be given in the one-shot manner. At this stage, percent of unit tests passed should count at least 10% of final score, percent of review tests passed at least 30%, and the final exam the balance. Note however, that students who have passed all unit and review tests have demonstrated acceptable achievement

in all portions of the course; therefore, in interpreting the results of the final exam such students' scores should not be viewed as less than satisfactory except in very unusual cases.

d. Wholely Repeated Tests Based Course. The last step in transitioning to repeated tests is to convert the final exam, i.e., the exam which covers the entire course content. This is appropriate only in cases where there is no need or requirement to make distinctions between the relative standing of various students, i.e., it is appropriate where the final course grade can be simply PASS/FAIL (or incomplete). In most educational settings there are sufficient reasons for wanting or needing more information than this, e.g., to award honors, scholarships, select the most capable for subsequent training, etc. Therefore, in general a one-shot test measuring how much students have achieved at the conclusion of a course should be continued.

2. Writing Equivalent, Alternate Test Forms.

Alternate forms of the same test (i.e., tests on the same unit or block of instruction) are prepared by using the same Test Plan Table. An effort should be made to construct the alternate forms for a particular unit such that they all have about the same level of difficulty. This will usually produce a set of alternate forms of a unit test that are sufficiently near equivalent for "classroom" testing purposes.

The actual degree to which alternate forms of a test approach true equivalence can only be determined retrospectively. If all forms were given to all students at one time, truely equivalent forms would all have about the same average score (mean), variability of scores (standard deviation) and scores on each form would be highly correlated with scores on each of the other forms.

In the initial phases of changing over to repeated tests, full use should be made of the best items available from tests used in the course at an earlier time. Note that since an "hour quiz" (i.e., review test) covers the same material as will be covered by several unit tests, items from an hour quiz used at an earlier time in the course may provide items for several unit tests. Also, because unit tests cover smaller amounts of material they need not be as long as the "hour quizzes".

3. Test Security.

Especially during the transition from one-shot testing to using a large number of alternate test forms, test security may be somewhat of a problem. In most cases, the practice of giving copies of tests to students after they have taken them will have to be discontinued. However, once the number of items used in the repeated test grows very large it will make little or no difference if the items are available to students just so long as they do not know which specific items will appear on the next alternate form of a test that they will take. This is true because when the number of items is quite large, memorizing answers to specific items either (a) is not as efficient as learning the underlying facts, principles, etc. which provide a basis for answering many questions or (b) such "memorization" produces exactly the same effect as learning the underlying facts, relationships, principles, etc.

Once a fairly substantial number of items covering a particular unit is available (i.e., at least two or three times the number of items that will be used in a single form of the unit test) multiple alternate forms can be easily generated by available computer programs. One such program, QUEST, allows you to have as many alternate forms as you wish drawn from an established item pool. Essentially, the program randomly selects items from the pool and prints them out on a master which can then be reproduced in sufficient quantities to meet the user's needs. (The UAMS Computing Facility is currently studying the need for making program QUEST available.)

4. Establishing A Criterion Score.

When repeated tests are used, a score must be chosen which distinguishes between those students with a satisfactory level of achievement (i.e., PASS) and those who need to learn more of the material (i.e., FAIL). This "cutting" score between pass and fail is called the criterion score.

The criterion score for most instructor prepared tests should usually be set between 70% and 100% correct. Very high criterion scores (i.e., 90%-100% correct) are appropriate only in cases where extremely limited, very well defined, and highly critical material is covered by the test, e.g., the signs and symptoms of cardiac arrest. This means very high criterion scores may be appropriate for some unit tests, but they will seldom be appropriate for a reivew test (because it covers the material in several units).

Where large quantities of information are covered and the test can only sample the material, a criterion of around 85% correct should be used. Empirical research has shown this to be the optimal value across a wide variety of subject matter contents and levels. However, 85% correct is a good criterion score only when the alternate forms have been contructed such that about 20% of the students in a typical class will attain criterion (i.e., pass the test) on their first attempt and sufficient time for re-study and alternate test forms for retesting are available for most students to eventually reach criterion. This effect is most easily achieved by setting the criterion (i.e., passing) score retrospectively rather than choosing the criterion score before the fact. For example, students could be told that the passing score for a unit test will be defined by that score immediately above which 20% of the students (in this class and/ or in previous classes) fall on their first attempt.

When very few alternate test forms are available, repeated testing is most effective in assuring only a "minimally satisfactory" level of achievement. When you have only two alternate forms of a unit test, the criterion should be set at that score above which about 80% of students fall on their first attempt (i.e., a score around 65-75% correct on a typical classroom test). Use of higher criterion scores (e.g., 85% correct) will effectively assure higher levels of achievement only when more alternate tests (and time for re-study) are provided.

The first time you use repeated tests within a course, determining the score above which a selected percentage of the students fall is facilitated by requiring all students to make their first attempt on a test for a given unit on the same date.

Also, if your class is large enough, about equal numbers of students should receive each of the alternate forms on which to make their first attempt. This will allow you to set different criterion scores for each alternate form in the event that different forms vary widely in difficulty. (Technically the best way to determine whether different cutting scores are needed is to use a one-way analysis of variance on the mean scores of the alternate forms. If the means are significantly different then seperate criterion scores are required. Otherwise, the same criterion score is used for all tests. However, "eye balling" the means and standard deviations of the tests is usually accurate enough to determine if the difficulties of the alternate forms are about the same.)

5. Scheduling Repeated Tests.

A schedule showing the date on which students must make their first attempt to pass each unit test (and review test where they are also given in the repeated tests mode) and the last date on which they may take an alternate test for each unit must be devised.

When repeated tests are first implemented in a course, all students should be required to make their first attempt on the same date. Generally this should be the day following the last formal presentation (i.e., lecture, seminar, laboratory, etc.) within a unit. The last day on which a student is permitted to make a subsequent attempt should be no later than the time at which half of the formal presentations for the next unit have been delivered. For example, if the first unit covered the four lectures in the first two weeks of a course, the students would be required to make their first attempt on the first unit test around the second Friday in the course. If the second unit in the course covered material presented over a three week period (i.e., third through fifth week of the course) those not passing on their first attempt on unit one would be permitted to continue taking alternate forms of the unit one test until whichever of the following occurred first (a) they had exhausted the available alternate test forms without passing, (b) they passed an alternate, or (c) Wednesday of the 4th week of the course passes. The specific schedule adopted in a given course should, on the one hand, discourage student procrastination; this is accomplished by fixing a definite last date on which an attempt can be made. On the other hand, a sufficient time interval between when a student may (or must) make his first attempt and when he can make his final attempt should be provided. In this way students are not discouraged from restudying the material prior to making an additional attempt. To assure maximum learning and to reduce the transitory effects of having taken a similar test before, at least 24 hours should be required to elapse between each attempt a student is allowed to make on a test (i.e., an alternate form of a test covering the same material). This twenty-four hour delay encourages students to actually re-study material before making another attempt to pass the test.

6. Administration of Repeated Tests.

The giving of repeated tests and the necessary record keeping associated with them need not be a burden on the faculty. Repeated unit tests can be given with minimal or no supervision. This is safe so long as (a) at least the "final exam" (i.e., review test covering the entire course which is given at the conclusion of the course) is given under conditions which permit reasonable assurance

that scores on the "final" represent each student's own work and (b) the final exam (combined with any other tests given under reasonable "supervision") has a weight of at least 30% in determining a student's overall final course grade. Under these conditions unit tests can be essentially self administered. One need only arrange for it to be inconvenient for students to copy specific items from the tests.

A clerk or secretary can maintain a file of multiple copies of each alternate form of each unit test, along with a schedule for when various tests may be taken. Then a student need only obtain a copy of a test he wants to take from this clerk. The clerk determines that the student is eligible for the test he has requested, provides the student with a copy of one of the forms that the student has not had before, if desired (by the instructor) the clerk "proctors" the test while the student takes it, the clerk or the student under the clerk's supervision hand scores the test, the clerk enters the numeric score in a record, determines whether it is above or below the criterion, enters "P" or "F" in the record for that student on that date for that form of that unit test, and informs the student whether he passed or failed. If the student has any question about the correct answers in the clerk's key, he is referred to a faculty member to resolve the problem.

The score and pass/fail entered by the clerk should be provisional if the student's answers are recorded on a machine scorable answer sheet. In this case, the clerk periodically takes the answer sheets to the Computing Facility to be machine scored. The score obtained through machine scoring is the official score. When this procedure is used, the machine scoring should be done early enough so that should a student have been told that he passed a test on the basis of the hand scoring but in fact there was an error in scoring and he failed the test, then he will have time to make one more attempt. This system is most efficiently implemented by providing a short period following the closing date for taking a particular unit test during which only those students who (a) were told in error that they passed their last attempt when they actually failed, and (b) have not made as many attempts as there are alternate forms of the test are allowed to make one further attempt to pass the unit test.

V. SCORING AND ANALYZING THE TEST

The purpose of scoring and analyzing the test is to provide the instructor useful information regarding how (a) students, (b) individual test items, and (c) the test as a whole performed. Even if all the guidelines given above were strictly followed, there would be no certainty that a test would work in the manner desired by the instructor. Following the above guidelines will greatly raise the probability that the test will function as desired, but until it is given to at least one group of students and the results properly analyzed, there is no way to determine how good a test it really is. Because of this, a test analysis is needed before student scores can be meaningfully interpreted with any real confidence. A test analysis should always be conducted whether the test is manually or machine scored. Where machine scoring is used, at least part of the test analysis work is usually done along with the test scoring. (This is the case with the machine test scoring service available through the UAMS Computing Facility.)

In the following sections, the procedures for computing two useful item statistics (i.e., Difficulty Index and Discrimination Index), and their uses in analyzing the test are described. These and other statistics discussed below are usually most meaningful when they are computed on the basis of the performance of students who are making their first attempt on a test which covers a particular block of instruction. Thus, when you use the Repeated Tests approach (see section IV. B.), the student data to analyze is that which comes from students making their first attempt on any of the alternate forms of a particular unit test. For example, suppose you have three alternate forms X, Y, and Z of the test for unit III of your course. Each of the alternate tests, i.e. III-X, III-Y, and III-Z should be analyzed separately. In analyzing test III-X, use the data from students for which their performance on this test represents their first attempt on any unit III test.

A. Item Difficulty.

The most common measure of item difficulty is the Difficulty Index. It should be computed for each item in the test. The Difficulty Index always has a value between 0.00 and 1.00. In some measurement and statistics texts, it is defined in such a way that bigger numbers mean that the item was easier. This is counterintuitive to many instructors. They argue that if the Difficulty Index is a measure of how hard (difficult) an item is, then bigger numbers should mean the item is more difficult; otherwise, it should be called the Item Ease Index. In recognition of the intuitive appeal of this argument, the Difficulty Index is defined below such that the bigger its value the harder the item was for the particular group on whose test data it was calculated. (This

definition is consistent with the one used in the current UAMS machine test scoring program, i.e., OPSCAN, except OPSCAN expresses Difficulty Indexes as Percentages instead of Proportions.) Although your numerical habits of a lifetime may incline you to think that a Difficulty Index of 0.80 indicates an item which is twice as hard as an item having a Difficulty Index of 0.40, this is not the case. Nor is it the case that the difference between Difficulty Index values of 0.30 and 0.50 means the same as the difference between 0.70 and 0.90. Difficulty Index values (as defined here) provide only rank order information. Thus, given the Difficulty Index values for several items you can arrange them in order of difficulty. But, the Difficulty Index does not provide sufficient information to make direct statements about the differences or ratio's of difficulties among items. Making comparative statements about the difference or ratio in difficulty between items requires specifying (or assuming) some kind of systematic relationship between the distribution of competence/skill among students and students' probability of providing the correct answer. Such considerations fall beyond the scope of this paper. (See Anastasi, 1968, p.166.)

1. Calculating the Difficulty Index.

The Difficulty Index should be calculated for each item in the test. It can be calculated only after you have first scored the responses of all the students to a given item. Once this is done, the appropriate computational procedure depends upon whether the item was scored on a Right-Wrong or a Partial Credit basis. The procedure given below for Partial Credit Scoring can be applied to Right-Wrong items as it is the more general of the two procedures. The procedure for Right-Wrong scoring is simply a special case but it allows you to reduce the amount of arithmetic involved. If the more general procedure (i.e., for Partial Credit Scoring) is applied to an item which is actually scored Right-Wrong, it will give exactly the same value for the item's Difficulty Index as would be obtained by using the simpler procedure intended for items scored Right-Wrong. (At present, the UAMS test scoring computer program computes Difficulty Indices only for items scored Right-Wrong.)

a. Right-Wrong Scoring. If you use a procedure to score an item which results in either all or none of the points associated with an item being given to a student (on the basis of his answer's correspondence to the correct answer), then your scoring procedure is basically Right-Wrong Scoring. For each item in the test which is scored in this manner, use the

following procedure to compute its Difficulty Index:

(i) Count the number of students who got the item wrong (i.e., failed to provide the correct answer). Omissions (i.e., failure to respond to an item) should be counted as wrong answers.

(ii) Divide the number of students who failed the item (NF) by the total number of students in the group(N). Round-off to two decimal places.

(iii) The resultant number is the Difficulty Index for this item. The above computation is summarized in the formula below:

$$\text{Difficulty Index} = \frac{NF}{N}$$

Where N = total number of students in the group; and
NF = the number of students who failed the item.

b. Partial Credit Scoring. If you use a test item scoring procedure which allows assigning something between all or none of the points associated with an item (on the basis of the degree of correctness of the student's answer), then your scoring procedure is of the Partial Credit variety. For each item in the test which is scored in this manner, use the following procedure to compute its Difficulty Index.

(i) Find the sum of the points made by all students on this item (ΣX);

(ii) Determine the total possible number of points (NV) this group of students could have obtained by multiplying the item's point value (V) by the number of students in the group (N);

(iii) Subtract the number of points made (ΣX) from the number of possible points (NV). (This intermediate result is the number of points lost and is analogous to the number of students who failed an item when Right-Wrong scoring is used.)

(iv) Divide the number of points lost (i.e., result of step iii above) by the possible number of points (NV). Round off to two decimal places. The result obtained is the Difficulty Index for this item, i.e., the proportion of points lost out of possible points that might have been lost. The above computational procedure is summarized in the formula below:

$$\text{DIFFICULTY INDEX} = \frac{NV - \sum_{i=1}^{N} Xi}{NV}$$

Where N = total number of students taking the test.
V = point value of item
Xi = score, i.e., number of points obtained by the ith student on this item.
i = index which identifies (counts) students, It runs from 1 through the value of N.

$\sum_{i=1}^{N}$ = arithmetic operator which means you are to find the sum of the ith through Nth value of the variable immediately to the right.

2. Interpreting the Difficulty Index.

Proper use and interpretation of the Difficulty Index depends upon whether the item is intended to measure specific, highly critical material or if the item is intended to only "survey" a general area, that is the item is to be only representative of many questions which it might be useful for a student to be able to answer about a general category or area of material. In either case, the guidelines provided (below) on item difficulty should be used in conjunction with the item Discrimination Index guidelines provided in section V. B. Also see Table 4.

a. Critical Material Items. If an item tests student mastery of highly critical, specific material, that is, some essential performance, the difficulty index indicates primarily (a) how successful you were in communicating the criticality of the material and (b) how effective your instruction and the students' study was in producing mastery of the material. Secondarily, it may suggest that the item was poorly stated. Whenever possible, mastery of truly critical material should be assessed using exhaustive tests, that is, tests in which every critical item is included rather than only a "representative sample". In addition, exhaustive tests should be given using the Repeated Tests approach. If the material is truly critical, anything less than a high level of mastery is not acceptable. Thus, for such material it is

clearly necessary for students to study and if necessary re-study the material until they can pass a test upon it at a high level of proficiency.

(i) Difficulty Index between 0.0 and 0.20. When the Difficulty Index is in this range for an item, you may conclude that the instruction has been quite effective.

(ii) Difficulty Index between 0.20 and 0.80. Instruction and study efforts by students have been only moderately effective. More instruction and study are required on the material covered by this item.

(iii) Difficulty Index between 0.80 and 1.00. The instruction and/or students' study efforts have been quite ineffective or the item was poorly written. If the item is faulty, it should not be counted in computing the scores of students. If the item is not faulty, revison in the instruction covering the material tested is indicated. The "revision" may take any or all of the following forms (a) more clearly communicating the importance of the material, (b) allowing more instructional and/or study time (c) using alternate and/or additional instructional methods (e.g., adding programmed instruction to lecture).

b. Representative Material Items. If an item is intended to be only representative of a large number of items that might be used to measure a student's competence in a general area; that is, if the item is used as a sample to survey a general area or large collection of material, the Difficulty Index should be interpreted according to the guidelines given below.

(i) Difficulty Index between 0.00 and 0.19. When the Difficulty Index is in this range it indicates that the question was really "too easy" for this group of students. This may arise because (a) most students already knew the answer prior to instruction (b) faulty item construction gave the answer away and/or (c) more emphasis was given (during instruction) to the material covered by this item than it deserved. It is highly probable that essentially all the students who missed

this item did so because of random errors such as mismarking their answer sheet or misreading the item, etc. In almost all cases "representative" items in the Difficulty range of 0.00 through 0.09 should be assumed to be "faulty". (See section V.C on scoring faulty items.)

(ii) Difficulty Index between 0.20 and 0.79. Representative items intended to survey a general area which have Difficulty Index values between 0.20 and 0.79 can usually be assumed to be good items. Thus, such items should be used either primarily or exclusively as the basis for determining scores on the test. Such items should be retained for use on tests with later groups of students.

(iii) Difficulty Index between 0.80 and 1.00. When the Difficulty Index is in this range, the item should usually be considered "too hard" for this group of students. This may arise because of (a) inadequate or faulty instruction on this portion of the material such as giving substantially less emphasis to the material than it deserved or (b) faulty item construction which led even the better students to answer incorrectly. It is highly probable that all or most of the (very few) students who answered this item correctly did so because of "random errors", such as guessing, mismarking the answer sheet, or misreading the item. In virtually all cases, "representative" items which have Difficulty Index values in the range of 0.90 through 1.00 should be assumed to be "faulty". (See section V.C. on scoring faulty items).

(iv) More Exact Guidelines For Difficulty Index of Good Items. Although the general rules of thumb provided above will usually suffice, the Difficulty Index values of good items can be more precisely defined. The difficulty of the item should fall mid-way between the level of performance expected if students were just guessing and that expected if all students knew the correct answer (i.e., Difficulty Index equal to 0.00). Table 2 provides this information for several types of items. Across all items in a test which are of

TABLE 2

IDEAL DIFFICULTY INDEX VALUE FOR VARIOUS ITEM TYPES

ITEM TYPE	IDEAL DIFFICULTY INDEX*
Two Option Multiple Choice (T-F)	0.25
Three Option Multiple Choice	0.33
Four Option Multiple Choice	0.38
Five Option Multiple Choice	0.40
Short Answer (Constructed Response)	0.50

*Note--Difficulty Index equals the proportion of students that fail an item (or alternately the proportion of total points lost on an item out of possible points that could have been made). The tabled values apply only in cases where each item is scored Right or Wrong. The values do not apply when partial credit scoring is used.

TABLE 3

ACCEPTABLE RANGE OF DIFFICULTY
INDEX VALUES FOR VARIOUS
ITEM TYPES

ITEM TYPE	ACCEPTABLE DIFFICULTY INDEX RANGE*
Two Option Multiple Choice (T-F)	0.12 through 0.38
Three Option Multiple Choice	0.17 through 0.50
Four Option Multiple Choice	0.19 through 0.57
Five Option Multiple Choice	0.20 through 0.60
Short Answer (Constructed Response)	0.25 through 0.75

*Note--Difficulty Index is defined as the proportion of students that fail an item (or proportion of points lost on an item out of possible points that could have been made). The tabled values apply only in cases where each item is scored Right or Wrong. The values do not apply when partial credit scoring is used.

the types given in Table 2, the average Difficulty Index should be near the value given for that type of item. The average of the Difficulty Index values for all items is equal to the average percent correct score for all students. Also, the Difficulty Index of each individual item should be fairly close to the value given in the table. What constitutes "fairly close" is a function of both the size of the ideal Difficulty Index and the number of students upon whom the observed value was obtained. Generally however, if the Difficulty Index is within the range defined by the value given in Table 2 plus or minus one half that value, then the Difficulty Index is acceptable. Table 3 provides the acceptable range of Difficulty index values for different item types. If the Difficulty Index for any "representative item" falls outside the indicated range, that item should usually be considered faulty. (See section V.C. on scoring faulty items).

B. Item Consistency.

Almost without exception, an instructor is justified in assuming that there should be some consistency between the overall performance of students on a test and students' performance on individual items within that test. Thus, students who do well (i.e., score high) on the whole test should do better on any specific test item than those students who do poorly on the test as a whole.

A number of different statistics are used to measure the degree of consistency between overall test scores and the scores on an individual item. These include various correlation coefficients (e.g., Pearson product moment, biserial, coefficient alpha, etc.) and the Discrimination Index. All of these statistics are interpreted in the same general manner. A positive value indicates that (a) students who do well on the test as a whole tend to do well on the item and (b) students who do poorly on the test as a whole tend to do poorly on the item. A negative value means the opposite (and inconsistent) relationship exists. That is, a negative value means that students who do well on the test as a whole tend to do poorly on the item. Values near zero indicate that there is no systematic relationship between performance on the item and performance on the test as a whole.

Because the Discrimination Index is (a) most easily computed, (b) easier to understand, and (c) provides as useful information as other more complicated item statistics, it is the only measure of consistency between an item score and total score that will be considered in detail here. (At present, the UAMS Test scoring computer program computes Discrimination Indices only for items scored Right-Wrong.)

1. Calculating the Discrimination Index.

The Discrimination Index should be calculated for each item in the test. Like the Difficulty Index, the Discrimination Index can be calculated only after you have first scored all the students' responses. Once this is done, the appropriate computational procedure depends on whether the item was scored Right-Wrong or on a Partial Credit basis. As with the Difficulty Index, if you are using the Repeated Tests approach, the Discrimination Index should be calculated from the data of students who are making their first attempt on any of the alternate forms of the test for a specific unit or block of instruction. Compute the Discrimination Index of each item in the test as follows:

a. Score each students test (i.e., all items).

b. Compute each students total score (i.e., the sum of the number of points earned).

c. Place the students' tests in rank order beginning with the highest total test score and ending with the lowest total test score.

d. Select a High and Low total test score group. Each group should contain an equal number of students. You may use the top and bottom 1/2, 1/3 or 1/4. Usually you get the clearest results when you approximate the upper and lower 30% of students. That is, take the 30% of student test papers with the highest total test scores and place them in the "High" group. Take the 30% of student test papers with the lowest total test scores and place them in the "Low" group. The remaining 40% (i.e., the middle group) can be disregarded during the computation of the Discrimination Index for each item on the test. If the class is relatively small you may need to use somewhat larger percentages to define your upper and lower groups. If possible, you should have at least 10 students' test papers in the High group and 10 in the Low group. The Discrimination Index becomes nearly meaningless with less than 5 students in both the High and Low groups. That is, the Discrimination Index becomes very unstable when computed for classes containing less than 10 students.

e. For each item, calculate the Difficulty Index for the High group alone using the appropriate procedure depending upon whether the item was scored on a Right-Wrong or Partial Credit basis. (See section V.A.1).

f. For each item, calculate the Difficulty Index for the Low group alone using the appropriate procedure depending upon whether the item was scored on a Right-Wrong or Partial Credit basis.

g. Subtract the Difficulty Index for the item calculated from the High group data from the Difficulty Index on the same item calculated from the Low group data.

Note that if the item was relatively more easy for the High group than the Low group (as it should be), this will produce a positive number. If all the High group got the item right, and all the Low group got it wrong your computation would be as follows:

Difficulty Index for Low Group	=	1.00
Difficulty Index for High Group	=	-0.00
Discrimination Index for this item	=	1.00

In this example the proportion of students who did well on the test as a whole (i.e., those in the High group) who found this item hard was smaller than the proportion of students who did poorly on the test as a whole (i.e., those in the Low group) who found this item hard. This produces a positive Discrimination Index for this item.

2. Interpreting the Discrimination Index

As with the Difficulty Index (as defined above) the Discrimination Index can only be interpreted in rank order terms. The Discrimination Index is sufficient for you to place items in order from those that are most discriminating to those that are least discriminating. But, Discrimination Index values cannot be properly interpreted in terms of ratios or interval differences in item discrimination. The interval difference between 0.10 and 0.20 does not equal the differnce between 0.50 and 0.60. Nor is an item with a Discrimination Index of 0.60 twice as discriminating as an item having a Discrimination Index of 0.30.

TABLE 4
ITEM QUALITY AS A FUNCTION
OF DIFFICULTY AND DISCRIMINATION

(A) Items testing highly specific, critical material

DIFFICULTY INDEX	DISCRIMINATION INDEX			
	-1.00 to -0.11	-0.10 to -0.01	0.00 to +0.09	+0.10 to +1.00
0.90 to 1.00	Faulty	Faulty	Faulty	Doubtful
0.80 to 0.89	Faulty	Faulty	Doubtful	Doubtful
0.20 to 0.79	Faulty	Faulty	Doubtful	Acceptable
0.10 to 0.19	Faulty	Doubtful	Doubtful	Acceptable
0.00 to 0.09	Faulty	Doubtful	Acceptable	Acceptable

(B) Items testing general, representative material

DIFFICULTY INDEX	DISCRIMINATION INDEX			
	-1.00 to -0.01	0.00 to 0.09	+0.10 to +0.19	+0.20 to +1.00
0.90 to 1.00	Faulty	Faulty	Doubtful	Acceptable
0.80 to 0.89	Faulty	Doubtful	Doubtful	Acceptable
0.20 to 0.79	Faulty	Doubtful	Acceptable	Acceptable
0.10 to 0.19	Faulty	Doubtful	Doubtful	Acceptable
0.00 to 0.09	Faulty	Faulty	Doubtful	Acceptable

a. Critical Material Items. When an item is intended to test specific, critical material, an item may be considered acceptable if (a) the Difficulty Index is between 0.20 and 0.79 and the Discrimination Index is equal to or higher than +0.10; or , (b) the Difficulty Index is equal to or below 0.19 and the Discrimination Index is not lower (i.e., more negative) than 0.00 Items measuring specific critical material which (a) have Difficulty Indexes between 0.20 and 0.79 and Discrimination Indexes below +0.00 (i.e., negative) should be considered faulty. See Table 4(A). (See section V.C. on scoring faulty items).

b. Representative Material Items. When an item is intended to be representative of a relatively large category of material, it should have a positive Discrimination Index (i.e., equal to or above 0.00) regardless of its Difficulty Index. When a representative item has a Difficulty Index value which falls either in the range of 0.00 to 0.19 or 0.80 to 1.00, it should have a Discrimination Index of at least +0.20. When the Difficutly Index is in the range of 0.20 to 0.79, the Discrimination Index should be at least 0.10. If any of the above conditions are not satisfied for a representative item it should be considered at least suspect and probably faulty. See Table 4 (B). (See section V.C. on scoring faulty items).

C. Scoring Faulty Items.

After scoring the test and performing the item analysis described above, you will almost always find some items in the test which the above guidelines suggest are technically faulty. In addition, when you conduct an after-test discussion of the test with your students, frequently a student will present a convincing argument why a response (i.e., answer) which you scored as wrong is as good as the response(s) you scored as right. As indicated above, you should have specified a procedure for identifying such faulty items and how they would affect student scores (and also communicated this procedure to students) prior to the students taking the test.

Whatever the specifics of the procedure you adopt, it should have the effect of eliminating the faulty items from your process of evaluating students' over-all test scores. That is, the procedure you use to eliminate faulty items should result in their having no impact upon the meaning (e.g., PASS/FAIL, A, B, C, etc.) you attributed to the total test score based on the good items. Basically, there are four ways to accomplish this effect and your procedure should conform to one or a combination of them.

1. Invalid (No Credit) Items.

Rescore each student's test giving no credit for any faulty item; that is, rescore the test as if the faulty items were just not there. This procedure is most appropriate where you have used Right-Wrong scoring and have pre-set cutting scores for specific grades (e.g., PASS/FAIL or A, B, C, etc). Notice that if you have pre-specified grades in terms of number of points earned or items passed, these will have to be adjusted down in an amount equal to the number of eliminated items (or their total point value). If you have pre-specified percentage cutting points for specific grades no adjustment need be made, because the student's official score will be based on the percentage of good items (i.e., those not deleted) that he answered correctly. You should be aware that this procedure may result in some student resentment if it is not explained prior to giving the test. A student who provides the "correct" answer to a faulty item, even if (as is frequently the case) he did so by guessing, is not going to appreciate having the points associated with that item "taken away from him". However, if you properly explain your procedure before hand, making clear that the reason for its use is to assure that faulty test items do not influence student grades, you will encounter very little student resentment or dissatisfaction.

2. Bonus Credit Items.

Rescore each student's test giving full credit to all students regardless of how or even whether they answered the faulty item(s). This procedure is most appropriate where you have used Right-Wrong scoring and you have not established preset cutting scores for specific grades (i.e., PASS/FAIL, or A, B, C, etc.) in absolute terms such as number or percent of items correct or points earned. Notice that you could use this approach where pre-set cutting scores were established for specific letter grades but the cutting scores would have to be adjusted upward in a amount equal to the number or point value of faulty items for which Bonus Credit was given. Students are likely to view this as "giving with one hand and taking away with the other". Note that if your "pre-set cutting points" are defined relativistically such as (a) the top 80% of students PASS, or (b) the top 10% of students get letter grade "A", or students attaining a standard score of 600 (i.e., one standard deviation above the mean; where, the mean equals 500 and the standard deviation equals 100) through 650 get letter grade "B",

then faulty items could be treated as either No Credit or Bonus Credit items. However, if a relativistic set of pre-specified letter grade standards are used, generally the Bonus Credit procedure is most "psychologically" acceptable to those students who one way or another managed to provide the "correct" answer to a faulty item.

3. Combination of Invalid and Bonus Items.

When the Difficulty Index is either very high or very low, the Discrimination Index is forced to be near zero. It is sometimes most intuitively meaningful under these circumstances to choose either the Invalid or Bonus procedure on the basis of how many students answered the question correctly. If 50% or less of the students answered the faulty item correctly, treat it as an Invalid (no credit) item; otherwise, treat it as a Bonus Credit item.

4. Partial Credit Items.

Change the scoring key for the item to include the additional "correct" answers and then rescore every student's test using a partial credit scoring procedure for the "faulty" item. This approach should be used only with short answer items and multiple-choice multiple-possible-correct-option items. It is not appropraite for T-F or simple multiple-choice (i.e., where the student is directed to select the single best or correct option).

5. Limitations on Deletions of Faulty Items.

In general, you should not "delete" more than about 20% of the items from a test even if more than that are technically faulty according to the Item Difficulty and Discrimination guidelines given above. You should delete first those items with the most negative Discrimination Indexes, then those with extreme Difficulty Indexes. Having done this the overall reliability of the test (i.e., based on only the remaining "good" items) should be high enough for the total test scores and grades derived from them to be meaningful. If this does not happen then you can usually conclude that there was something seriously amiss either in the manner in which the test was constructed or in the instruction provided on the material covered by the test. If the guidelines for building and giving tests given here are followed, it should be exceedingly rare that more than 20% of the items in any test you build and give will be in serious violation of Item Difficulty and Discrimination guidelines suggested for good items.

D. Test Reliability.

Test reliability concerns the overall performance of the whole test. It is of most importance when test scores are used to make comparisons among various students who took the test. Reliability is concerned with the accuracy and dependability with which test scores actually reflect true differences between those who made high scores and those who made relatively lower scores. If the high scorers are really the better students, then another equivalent test should result in scores which arrange students in much the same order as the first test. Thus, grades (e.g., PASS/FAIL, or A, B, C) that were assigned on the basis of a highly reliable test would not be changed much ,if at all, were you to give another equivalent, highly reliable test. One of the things which contributes most to the reliability of a test is the consistency of the items within the test. For this reason, elimination of items having negative Discrimination Index values almost invariably raises the reliability of the test based on the remaining items. However, reliability also depends on the length of the test (i.e., the number of items it contains). A test twice as long as another which has an equal quality of items will be somewhat more reliable (although generally much less than twice as reliable). If there is so little consistency among the items on a test that deletion of the most inconsistent 20% of the items fails to produce an adequate reliability, there are not likely to be enough items remaining that a further reduction would yield a test containing a sufficient number of consistent items to make a reliable test.

There are a number of different procedures for determining the reliability of a test. However, all the "exact" procedures involve rather long and tedious computations. Considering the work load of most instructors, use of these procedures hardly seems time effective. Where machine scoring is used, the computer program used to score the test and generate item statistics will usually provide a test reliability coefficient as well. (The UAMS' machine test scoring program provides this information; however, at present the reliability coefficient is based only upon items which are scored Right-Wrong).

1. Appropriate Reliability Coefficient Values.

For short, less critical tests (e.g., hour quizzes) a reliability of at least 0.60 should be obtained if the instructor is to have any confidence in the meaningfulness of the scores on the tests. Reliabilities in the range of 0.70 to 0.80 should be considered good. Those above 0.80 are quite good for "classroom" tests. For highly important tests (e.g., final course examination, qualifying exam, or admissions tests) reliabilities below 0.70 are not acceptable. These higher reliabilities

are most usually achieved through the simple expedient of making really important tests fairly long. Final exams containing 300 or more items are required in many content areas to achieve reliabilities in excess of 0.80. However, high reliability can be achieved in relatively short tests by building them from items which were used in previous tests and which had good characteristics (i.e., moderate difficulty and positive discrimination).

2. Reliability versus Validity.

A note of caution about test reliability is needed. When tests are of the "mastery" variety, that is when virtually all students will make very nearly the same (high) score on the test, the accepted procedures (and the conceptual rationale justifying them) for calculating a reliability coefficient do not apply. To obtain reasonably high reliability coefficients the scores of students must be spread out. On a mastery test this will just not happen except on its first administration (i.e., while there is still a large difference among the students with regards to how well they have mastered the material covered by the test). When the repeated tests (i.e., mastery) approach is used the reliability coefficient is really meaningful only on the students' first attempt to pass any of the alternate forms of the unit test. Even when the one-shot testing approach is used, the reliability coefficient should not be viewed as the sole or even most important guide to the worth of the test. Proper coverage of relevant material through the use of a test plan (or equivalent method) is one fairly sure way to insure that the test is measuring what you want it to measure (i.e., that the test is valid). If the test is not valid, it makes very little difference whether or not it has high reliability. For this reason, the above guidelines first emphasized proper construction of the test to cover the material you want it to cover. Once this is done, you should attend to the adequacy of the test's reliability. If deletion of faulty items does not bring the reliability into the acceptable range, then you simply need to give another test. Treat this added test as just that many items added to the first test. It is always possible (although sometimes quite time consuming) to "lengthen" the test until it becomes both adequately reliable and valid. A few minutes study of Tables 5 and 6 should help clarify the relationships between reliability, test length, test difficulty, and variability (i.e.,standard deviation) of test scores.

3. Finding the Approximate Reliability of a Test.

Because the "exact" formulas for estimating the reliability of a test involve rather time consuming computation, two short-cut methods for getting a rough approximation are given below. One requires doing a little arithmetic to estimate the Standard Deviation of the test. If even this seems too much, you can use the second method where you simply assume that the standard deviation is about 0.15 times the number of items in the test (or total point value of all items in the test).

a. Estimating the Test's Standard Deviation. Arrange the test papers in order; highest to lowest score. Write the scores on a sheet of paper. Take the top one-sixth of the students' scores and find their sum. If one-sixth of the students do not turn out to be a whole number of students, then add in the fraction of the next students score that was "left over". For example, if you had 40 students then 40 divided by 6 = 6.33. Find the sum of the test scores for the top 6 students. To this add 0.33 times the seventh student's score. Now do this for the bottom one-sixth of the students. Subtract the sum for the bottom one-sixth from the sum for the top one-sixth. Divide the result by one-half the number of students in the whole class. Round off the results of the division to the nearest whole number. This gives you the Standard Deviation of the test to the nearest whole number. This procedure is summarized in the formula below:

$$\text{Standard Deviation} = \frac{\text{Sum of top 1/6th - sum of bottom 1/6th}}{\text{Half the number students}}$$

Unless the distribution of test scores departs drastically from a "bell-shaped curve" (normal distribution), this procedure will give the same value as more conventional methods of calculating the standard deviation of the test.

b. Method 1 for Approximating the Test's Reliability. Use this method when you have estimated the Standard Deviation as suggested above. If the average (mean) score on the test is between 70% and 90% correct, use Table 5. If the average score is between 50% and 70%, use Table 6. Divide the Standard Deviation (as estimated from the procedure given above) by the number of items. If the result is close to 0.10, use the top line in the Table. If the result is closest to 0.15 use the second line in the table. Then

choose the column that is nearest to the number of items in your test. The figure (number) at the intersection of the selected row and column is the approximate reliability of your test.

c. Method 2 for Approximating the Tests' Reliability. Use this method if its just too much trouble to first estimate the test's standard deviation as suggested in section D.3.a above. Assume that the standard deviation divided by the number of items on the test is around 0.15 (as is frequently the case). Use line two in the appropriate table (5 or 6) and proceed as in Method 1 above.

Table 5*
Approximate Reliability of "Easy" Tests (average 70 to 90% correct)

Number of Items (N)	20	30	40	50	60	70	80	90	100
SD=0.10N	.21	.48	.62	.69	.75	.78	.81	.83	.85
SD=0.15N	.68	.80	.84	.88	.90	.91	.92	.93	.94
SD=0.20N	.84	.90	.92	.94	.95	.96	.96	.97	.97

SD=Standard Deviation

Table 6*
Approximate Reliability of "Hard" Tests (average 50 to 70% correct)

Number of Items(N)	20	30	40	50	60	70	80	90	100
SD=0.10N	-	.21	.41	.53	.61	.66	.71	.74	.77
SD=0.15N	.49	.67	.75	.80	.84	.86	.88	.89	.90
SD=0.20N	.74	.83	.87	.90	.92	.93	.94	.94	.95

*Tables 5 and 6 from Shortcut Statistics for Teacher-Made Tests (Educational Testing Service, 1964, page 31).

VI. INFORMING STUDENTS AFTER THE TEST

Following the scoring and analysis of the test, the following guidelines should be followed in providing information to students.

A. Information to be Provided.

Provide students with the following kinds of information.

1. Student's total test score and interpretation in terms of a letter grade or Pass/Fail where applicable.
2. Distribution of total test scores.
3. What items were identified as faulty, not used in computing total test scores, and why.

In addition each student should always be provided an opportunity to go over his own test and confirm that the scoring was done according to the "correct answer" key.

B. After Test Discussion Session.

An item by item review and discussion of the test should be provided to all students. Special emphasis should be placed upon any items which were identified as faulty on the basis of Item Difficulty Index or Discrimination Index. Students should be provided an opportunity to challenge the appropriateness of the "correct answer" key. When a student makes a convincing challenge, the item should be rescored for all students according to the guidelines for scoring faulty items. During the discussion session, students should be informed why "correct" answers are correct as well as why "incorrect" answers are incorrect for any item(s) in which the students are interested.

C. Confidentiality.

Information regarding the performance of individual students is confidential and prudent measures should be taken to safeguard it from release to unauthorized individuals or groups. Since the passage of the Federal legislation commonly called the "Privacy Act", maintaining the confidentiality of student records has been mandated by law. To protect yourself, you should be very circumspect about releasing information on individual students. However, this should not be taken to mean that you can withhold information from a student about himself. Under the provisions of the so called "Freedom of Information" act, an adult student is entitled to full disclosure of any information in any file which is about him. The exceptions to the full disclosure rule provided by law will very seldom (if ever) apply to the information in educational files.

Because of the tradition of "posting grades", a special word of caution seems appropriate. If you are going to "post" test scores or grades be sure to use some method that will allow only the student concerned to determine his own score or grade and that no one can determine the score or grade for another person.

BRIEF BIBLIOGRAPHY

Anastasi, A. Psychological testing. (Third Edition). New York, N.Y.: Macmillan, 1968.

Bloom, B.S. Taxonomy of educational objectives, handbook I: Cognitive domain. New York, N.Y.: McKay, 1956.

Deese, J. and Hulse, H.H. The psychology of learning. (Third Edition) New York, N.Y.: McGraw-Hill, 1967.

Ebel, R.L. Measuring educational achievement. Englewood Cliffs, N.J.: Prentice Hall, 1965.

Educational Testing Service. Making the classroom test: A guide for teachers. Princeton, N.J.: Educational Testing Service, 1961.

Educational Testing Service. Short-cut statistics for teacher made tests. Princeton, N.J.: Educational Testing Service, 1964.

Hilgard, E.R. and Bower, G.H. Theories of learning. (Third Edition) New York, N.Y.: Appleton-Century-Crofts, 1966.

*Hively, W. Domain-referenced testing. Englewood Cliffs, N.J.: Educational Technology Publication, 1974.

*Keller, F.S. and Sherman, J.G. The Keller plan handbook. Menlo Park, California: W.A. Benjamin, Inc., 1974.

*Lippey, G. (Ed.) Computer assisted test construction. Englewood Cliffs, N.J.: Educational Technology Publication, 1974.

Lord, F.M. The relation of reliability of mulitple-choice tests to the distribution of item difficulties. Psychometrica, 1952, 17, 181-194. (a)

Lord, F.M. A theory of test scores. Psychometric Monographs, 1952, No. 7. (b)

McNemar, Q. Psychological statistics. New York, N.Y.: Wiley, 1962.

*Mager, R.F. Preparing instructional objectives. Belmont, California: Flaron Publishers, 1962.

Nunnally, J.C. Psychometric theory. New York, N.Y.: McGraw-Hill, 1967.

*Popham, W.J. (Ed.) Criterion-referenced measurement. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

*Shoemaker, D.M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45(1), 127-147.

*Tuckman, B.W. Measuring educational outcomes: Fundamentals of testing. New York, N.Y.: Harcourt-Brace-Jovanovich, Inc., 1975.

*Available in the Office of Educational Development library: Room B-305 (basement of ED I).